Human DNA Sequence Variation in a 6.6-kb Region Containing the Melanocortin 1 Receptor Promoter

Kateryna D. Makova,* Michele Ramsay,[†] Trefor Jenkins[†] and Wen-Hsiung Li*

**Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and* [†]*Department of Human Genetics, South African Institute for Medical Research, Johannesburg, 2050 South Africa*

Manuscript received December 13, 2000
Accepted for publication April 19, 2001

ABSTRACT

An ∼6.6-kb region located upstream from the melanocortin 1 receptor (MC1R) gene and containing its promoter was sequenced in 54 humans (18 Africans, 18 Asians, and 18 Europeans) and in one chimpanzee, gorilla, and orangutan. Seventy-six polymorphic sites were found among the human sequences and the average nucleotide diversity (π) was 0.141%, one of the highest among all studies of nuclear sequence variation in humans. Opposite to the pattern observed in the MC1R coding region, in the present region π is highest in Africans (0.136%) compared to Asians (0.116%) and Europeans (0.122%). The distributions of π, θ, and Fu and Li's *F*-statistic are nonuniform along the sequence and among continents. The pattern of genetic variation is consistent with a population expansion in Africans. We also suggest a possible phase of population size reduction in non-Africans and purifying selection acting in the middle subregion and parts of the 5′ subregion in Africans. We hypothesize diversifying selection acting on some sites in the 5′ and 3′ subregions or in the MC1R coding region in Asians and Europeans, though we cannot reject the possibility of relaxation of functional constraints in the MC1R gene in Asians and Europeans. The mutation rate in the sequenced region is 1.65×10^{-9} per site per year. The age of the most recent common ancestor for this region is similar to that for the other long noncoding regions studied to date, providing evidence for ancient gene genealogies. Our population screening and phylogenetic footprinting suggest potentially important sites for the MC1R promoter function.

STUDIES of human genetic variation provide a powerful means for elucidating the genetic, evolutionary, and demographic factors shaping the human genome. Such studies have recently been greatly facilitated by the advent of fast sequencing techniques and the abundance of human genomic sequence data, thanks to the efforts of the Human Genome Project. Recent surveys of human genetic polymorphism at the DNA sequence level can be divided into two major groups. The first group investigated regions of the human genome with no known or predicted genes, *i.e.*, noncoding regions (*e.g.*, Kaessmann *et al.* 1999; Zhao *et al.* 2000), and provided baseline data for studying neutral evolution. The second group focused on regions containing genes of particular interest (*e.g.*, Nickerson *et al.* 1998, 2000; Rieder *et al.* 1999) and presented a genetic framework for association studies between genotype (or haplotype) and phenotype (usually disease related). Despite the rapid accumulation of data on human genetic polymorphism (reviewed in Przeworski *et al.* 2000 and Yu *et al.* 2001), however, it is apparent that many more genomic regions will need to be analyzed to understand the nature of human genetic variation in all its complexity.

In the present study we examined genetic variation in a 6.6-kb region (on chromosome 16) that is noncoding but is located immediately upstream from the coding region of a well-studied gene, the melanocortin 1 receptor (MC1R). Our purpose is threefold. First, MC1R is a key regulator of melanin synthesis (MC1R expression increases the eumelanin to phaeomelanin ratio in skin) and is the only gene identified thus far to contribute to normal skin pigmentation variation in humans (Valverde *et al.* 1995; Abdel-Malek *et al.* 1999). However, MC1R specifies only a part of the normal pigmentation spectrum. Some mutations in the MC1R coding region are associated with light skin and red hair. Indeed, the majority (but not all) of red-haired individuals have mutations in the MC1R coding region (Harding *et al.* 2000). Recent studies of genetic variation in the MC1R coding region in worldwide populations (Rana *et al.* 1999; Harding *et al.* 2000) revealed MC1R haplotypes shared among humans with very different pigmentation phenotypes, indicating that there are additional genetic determinants of skin color in humans. Before turning to other pigmentation loci, we examined whether genetic variation exists in the promoter region of MC1R among humans, because such variation may be partly responsible for human pigmentation spectrum. Second, the variation in the coding region of the MC1R gene shows unusual features. The MC1R coding region is highly polymorphic in Asians

Corresponding author: Wen-Hsiung Li, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th S., Chicago, IL 60637. E-mail: whli@uchicago.edu

and Europeans with a strong excess of nonsynonymous substitutions (Rana *et al.* 1999; Harding *et al.* 2000). On the other hand, the MC1R coding region is conserved among Africans, where mainly synonymous substitutions have been found (Rana *et al.* 1999; Harding *et al.* 2000; John and Ramsay 2000). This pattern contrasts with the data from other loci where Africans are usually the most polymorphic (reviewed in Przeworski *et al.* 2000; Yu *et al.* 2001; with the notable exception of the Duffy locus, Hamblin and Di Rienzo 2000). The data led to the conclusion of strong purifying selection at the MC1R coding region in Africans (Rana *et al.* 1999; Harding *et al.* 2000). However, the data showing the pattern of variation observed in European and Asian populations can be explained in two different ways, namely (1) selection for variants determining lighter skin (Rana *et al.* 1999) or (2) relaxation of functional constraints (Harding *et al.* 2000). By sequencing the promoter region of MC1R we aim to investigate which explanation is more plausible. Third, our purpose was to compare the pattern of variation and evolutionary parameters estimated from this genomic region with those of other regions studied by the same sampling scheme, namely by examination of worldwide collections of one to six individuals per population (Kaessmann *et al.* 1999; Zhao *et al.* 2000; Yu *et al.* 2001). This sampling design can cover a wider spectrum of allelic variants worldwide as compared to studying polymorphisms in several "reference" population samples (Przeworski *et al.* 2000).

The 6.6-kb region we studied has been partially characterized. It contains a 3.2-kb previously published sequence located upstream from the MC1R coding region (Moro *et al.* 1999; accession no. AB026663). Multiple transcription initiation sites were detected within the ~600-bp sequence immediately upstream from the start codon of MC1R. Binding of a basal transcription factor SP-1 at three sites within the 1200 bp upstream from the start codon was shown by gel shift assay (Moro *et al.* 1999). There are also consensus sites of regulatory elements AP-1, AP-2 (binding sites of activating proteins 1 and 2), two TATA-boxes (however, situated distantly from the transcription initiation sites), and several E-boxes (binding sites of basic/helix-loop-helix/leucine zipper transcription factors). A promoter assay using luciferase as a reporter gene revealed that the minimal region exhibiting promoter activity (later called the "minimal promoter") was located within 517 nucleotides upstream from the start codon (Moro *et al.* 1999). Importantly, the 3.2-kb region studied by Moro *et al.* (1999) does not contain a complete promoter, and a regulatory element that silences expression of MC1R in nonspecific tissues (other than skin) is likely to be present outside of this region but remains to be identified. In addition, another ~5 kb of noncoding sequence located farther upstream was available in GenBank (accession no. AC008145). A 3′ untranslated region (UTR) of the KIAA1049 protein gene, expressed in the brain (Kikuno *et al.* 1999), is located immediately upstream from this noncoding sequence.

We analyzed the 6.6-kb region in 54 humans (18 Africans, 18 Asians, and 18 Europeans) and three outgroups (chimpanzee, gorilla, and orangutan). Additional evolutionary comparisons were made with the mouse MC1R promoter region (Adachi *et al.* 2000). Specifically, we are interested in the following questions. (1) What is the pattern of variation in this promoter region compared to the noncoding regions studied with a similar sampling scheme? (2) What is the distribution of variation among individuals in different continents? (3) Is the variation evenly distributed throughout the sequenced region, including the MC1R minimal promoter and *Alu* sequences? (4) Are the data compatible with neutral evolution? (5) Are the mutation rate (μ), the effective population size (N_e), the population parameter $\theta = 4N_e\mu$, and the age of the most recent common ancestor (MRCA) of the sequences in a sample estimated from this region different from those of other regions? (6) Can we identify sites with potential regulatory function in the sequenced region from population screening and comparison with the outgroup primate and mouse sequences?

MATERIALS AND METHODS

DNA samples: DNA used for this study was from 18 Africans (5 Nigerians, 4 South African Bantu speakers, 2 Biaka pygmies, 2 Mbuti pygmies, 1 !Kung, 1 Kenyan, 1 Kikuyu, 1 Zulu, and 1 Ghanian), 18 Asians (6 Chinese, 5 Indians, 3 Japanese, 2 Vietnamese, and 2 Cambodians), and 18 Europeans (2 French, 2 Germans, 2 Russians, 2 Italians, 2 Swedes, 2 Ukrainians, 1 Finn, 1 Hungarian, 1 Spaniard, 1 Portugese, 1 Norwegian, and 1 Dutch-Irish). To obtain outgroup sequences we used DNA from the common chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and orangutan (*Pongo pygmaeus*).

PCR and sequencing: We sequenced a total of 6.6 kb of noncoding segments between 3′-UTR of the KIAA 1049 protein gene and the MC1R gene coding region (Figure 1). A minisatellite of ~1 kb in length and an ~0.3-kb region including parts of the two middle *Alu*S repeats were not sequenced. Both regions proved to be difficult templates. The distribution of repeats within the sequenced region is shown in Figure 1. There are five *Alu* repeats: four *Alu*S's located next to each other and one *Alu*Y.

A 6.6-kb noncoding region was amplified in five parts (sequences of the PCR and sequencing primers are presented in the appendix) by touchdown PCR (Don *et al.* 1991): two overlapping fragments covering positions 30307 to 27820 of GenBank contig AC008145 and three overlapping fragments covering positions 26657 to 23086 (the region between 25037 and 24784 was not sequenced for the reasons given above). The PCR conditions were as described in Zhao *et al.* (2000). The PCR products were isolated from agarose gels and purified with a gel purification kit (QIAGEN Inc., Valencia, CA). Sequencing primers were designed every 400–450 bp in both directions. Sequencing reactions were performed according to the protocol of the ABI Prism BigDye Terminator sequencing kit (Applied Biosystems, Foster City, CA) modified by quarter reaction. Sequencing reactions were purified by Seph-

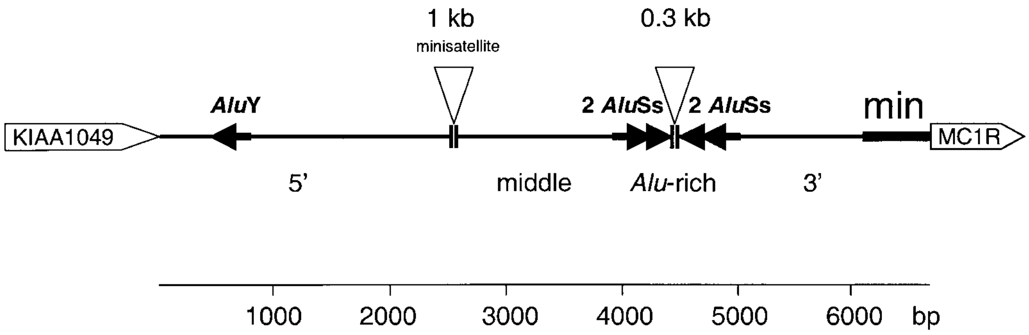

Figure 1.—Schematic representation of the sequenced region. Min, the minimal promoter and 5′-UTR of the MC1R gene. The 5′, middle, *Alu*-rich, and 3′ subregions are indicated. Arrows specify *Alu* repeats. Triangles indicate regions that were not sequenced. We kept the numbering contiguous in spite of two gaps in the sequence. The KIAA1049 and MC1R genes are not shown to scale.

adex G-50 (DNA grade, Pharmacia, Peapack, NJ) and run on an ABI 377 DNA sequencer using 4.25% gels (Sooner Scientific, Garvin, OK). Sequence in both directions was obtained for each amplified product. We did not determine the exact numbers of mononucleotides in poly(A) tails of *Alu* repeats.

ABI DNA sequence analysis 3.0 was used for lane tracking and base calling. The data were proofread manually and heterozygous sites were detected as double peaks. For each individual, sequences were assembled separately using SeqMan in DNAStar (DNAStar, Madison, WI). Assembled files were carefully checked by eye. Consensus sequences for all individuals were then aligned using MegAlign. Fluorescent traces for each variant site were rechecked again in all individuals. Additionally, all singleton, doubleton, and tripleton sites (variants that appear, respectively, only once, twice, or three times in the total sample) were verified by reamplification and resequencing. No errors were found. All sequences were submitted to GenBank under accession nos. AF387914–AF387969.

Statistical analyses: The more frequent nucleotide at each polymorphic site in the pooled sample of 54 sequences was selected for the human consensus sequence. The human ancestral sequence was inferred from comparison with the outgroup sequences using parsimony. Nucleotide diversity (π) and its standard error (derived from sampling variance) within and between continents were calculated using DnaSP ver. 3.50 (Rozas and Rozas 1999). The Watterson estimator θ and its standard error (derived from sampling variance assuming no recombination) per site were estimated from S (the total number of polymorphic sites) using DnaSP. The distributions of π and θ along the sequence were computed using the sliding window option of DnaSP ver. 3.50 with the window size of 750 bp and step size of 25 bp. The distribution of K_{JC}, the average number of nucleotide substitutions per site between species (human, chimpanzee, and gorilla) with the Jukes-Cantor correction, along the sequence was calculated with DnaSP using the same sliding window and step size as above. Repeats were identified with RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). F_{ST} was estimated according to Wright. Statistical significance of differences in allele frequencies at individual sites among the three continents was computed using a χ^2 test with Bonferroni correction for multiple tests.

The HKA test (Hudson *et al.* 1987) was performed using the manual mode of DnaSP and the divergence was calculated by comparison with the chimpanzee sequence. Tajima's (1989) test and Fu and Li's (1993) neutrality tests with an outgroup were performed using the program at Dr. Fu's website (http://hgc.sph.uth.tmc.edu/fu). The critical values for the significance of neutrality tests were obtained from 5000 simulated samples. The distribution of Fu and Li's F along the sequence was obtained using the sliding window option of DnaSP with the same window and step size as above.

The average numbers of nucleotide differences between human and outgroup sequences were calculated using DAMBE (Xia 2000). Pseudohaplotypes were first generated for each sequence. Gaps were deleted sitewise. The mutation rate per nucleotide per year (ν) was calculated according to $\nu = d/(2t)$, where d is the number of nucleotide substitutions per nucleotide site between two sequences and t is the divergence time between the two species. The mutation rate per sequence per generation was calculated as $\mu = \nu gL$, where L is the sequence length (bp) and g is the generation time (human $g = 20$ yr). Watterson's (1975) and Tajima's (1983) methods were used to estimate $\theta = 4N_e\mu$, where N_e is the effective population size.

The age of the most recent common ancestor of the sequences in a sample was calculated using Fu's (1996) and Fu and Li's (1996, 1997) methods. The mode, mean, and 95% confidence interval were computed in terms of years.

Potential binding sites of transcription factors in the human consensus and variant sequences were predicted using TRANSFAC (Wingender *et al.* 2000). An alignment with the mouse sequence was obtained using Advanced Pipmaker (http://bio.cse.psu.edu/pipmaker/) with the chaining option (Schwartz *et al.* 2000).

RESULTS

Pattern of sequence variation: We sequenced ~6660 bp of the selected region in 54 humans, one chimpanzee, and one gorilla, and 5789 bp in one orangutan. The GC contents of the sequences are ~59%, which is much higher than the genome average of ~42%.

A total of 76 variant sites were found among the 108 human chromosomes (Table 1). This included 72 nucleotide substitution sites (95%) and 4 insertions/deletions (indels; 5%). All 72 nucleotide substitution sites had only two alternative nucleotides: 58 (81%) were transitions and 14 (19%) were transversions. Among the four indels, there were two one-nucleotide indels and two two-nucleotide indels. On average 11 or 12 variant sites were found per 1000 bp within the studied region. Among the 76 variant sites (Table 1), 40 (including two indels) were mutations found only in one sequence in the sample (singletons), 6 (including one

TABLE 1

Numbers of variant sites (including indels) and nucleotide diversity (π) among the sequenced regions

Type of variant	6.6 kb in 16q24.3[a]			10 kb in 1q24[b]			10 kb in 22q11.2[c]			10 kb in Xq13.3[d]		
	All (108)[e]	Afr. (36)	Non-Afr. (72)	All (122)	Afr. (40)	Non-Afr. (82)	All (128)	Afr. (40)	Non-Afr. (88)	All (69)	Afr. (23)	Non-Afr. (46)
Singletons	40 (2)[f]	26 (1)	14 (1)	19 (3)	7 (3)	12 (0)	20 (2)	8 (1)	12 (1)	19 (0)	11 (0)	8 (0)
Doubletons	6 (1)[g]	4 (0)	1 (1)	7 (0)	2 (0)	5 (0)	23 (1)	19 (0)	4 (1)	5 (0)	5 (0)	0 (0)
Others	30 (1)	28 (1)	27 (1)	22 (1)	20 (1)	11 (0)	32 (1)	27 (1)	28 (0)	9 (0)	8 (0)	9 (0)
Total	76 (4)	58 (2)	42 (3)	48 (4)	29 (4)	28 (0)	75 (4)	54 (2)	44 (2)	33 (0)	24 (0)	17 (0)
π (%)	0.141	0.136	0.136	0.058	0.076	0.046	0.088	0.085	0.091	0.034	0.035	0.035

[a] This study.
[b] From Yu *et al.* (2001).
[c] From Zhao *et al.* (2000).
[d] From Kaessmann *et al.* (1999).
[e] No. of chromosomes studied.
[f] The number of indels is given in parentheses.
[g] One doubleton was shared between an African and an Asian.

indel) were found in two sequences (doubletons), and 30 (including one indel) were found in more than two sequences (others). There was an excess of low frequency variants (singletons and doubletons) compared to the other types of variants (46 *vs.* 30).

Two estimates of nucleotide variation were calculated (Table 2). The nucleotide diversity (π), *i.e.*, the average pairwise sequence difference between two random sequences in a sample, was 0.141% per site. The average estimate of θ, which is based on the observed number of polymorphic sites in a sample, was 0.211% per site. Under the neutral Wright-Fisher model, these two estimates should be equal. A higher value of θ than π (and consequently a negative Tajima's *D*) implies an excess of low frequency variants compared to high frequency variants, though the excess was not statistically significant (Table 5).

Distribution of diversity among continents: The numbers of variant sites in African, Asian, and European sequences were 58, 32, and 33, respectively. Indians were grouped with East Asians as "Asians" and not with the Europeans on the basis of geography and this grouping is supported by nucleotide diversity (π) between populations calculated from the data. The nucleotide diversity between the Indian and East Asian sequences studied (0.128%) was slightly lower than that between the Indian and European sequences (0.136%). Altogether, the 36 African sequences had 58 variant sites, whereas the 72 non-African sequences had only 42 variant sites (Table 1). There were 34 unique (not present in the other continents) variant sites (including 26 singletons)among the African sequences, while there were only 7 (all singletons) and 9 (7 singletons) unique variant sites among the Asian and European sequences, respectively. Thus, Africans had the largest proportion of variant sites (χ^2 = 10.63, d.f. = 2, $P < 0.005$) and the largest proportion of unique variants among variant sites (χ^2 = 34.14, d.f. = 2, $P < 0.001$).

In Africans, the number of low frequency variants (Table 1) was only slightly higher than the number of other variants (30 *vs.* 28), whereas in non-Africans there were almost 50% fewer low frequency variants than the other variants (15 *vs.* 27), resulting in a positive though nonsignificant Tajima's *D* (Table 5). The proportion of singletons was higher in Africans (26 singletons out of 58 polymorphic sites) than in non-Africans (14 singletons out of 42 polymorphic sites), but the difference was not statistically significant (χ^2 = 1.35, d.f. = 1, P = 0.25).

The average pairwise nucleotide diversity (π) was highest in Africans (0.136; 95% C.I. = 0.114–0.158), intermediate in Europeans (0.122; 95% C.I. = 0.104–0.140), and lowest in Asians (0.116; 95% C.I. = 0.094–0.138; Table 2), though all differences were not statistically significant. The π-value in non-Africans (when Asians and Europeans were considered together) was equal to that in Africans (0.136%). In contrast, the θ-value (Table 2) was almost twice as high in Africans

TABLE 2

Distribution of nucleotide diversity and Tajima's test statistic along the sequence

	Africans		Asians		Europeans		Total	
Sequence	π	θ	π	θ	π	θ	π	θ
5′ subregion	0.098	0.175	0.073	0.066	0.089	0.066	0.094	0.156
(1–432, 715–2488[a])	0.072–0.124	0.043–0.307	0.053–0.093	0.002–0.130	0.075–0.103	0.002–0.130	0.082–0.106	0.054–0.258
w/o *AluY*	$D = -1.45$ ($P > 0.1$)		$D = 0.30$ ($P > 0.1$)		$D = 0.98$ ($P > 0.1$)		$D = -0.90$ ($P > 0.1$)	
AluY	0.393	0.682	0.380	0.596	0.306	0.511	0.386	0.874
(433–714)	0.283–0.503	0.072–1.292	0.254–0.506	0.040–1.152	0.196–0.416	0.461–0.561	0.318–0.454	0.248–1.500
	$D = -1.24$ ($P > 0.1$)		$D = -1.03$ ($P > 0.1$)		$D = -1.10$ ($P > 0.1$)		$D = -1.49$ ($P > 0.1$)	
Middle subregion	0.079	0.182	0.076	0.091	0.082	0.109	0.086	0.215
(2489–3819)	0.049–0.109	0.028–0.336	0.052–0.100	0–0.185	0.056–0.108	0.003–0.215	0.070–0.102	0.067–0.363
w/o *Alu*	$D = -1.72$ ($0.1 > P > 0.05$)		$D = -0.42$ ($P > 0.1$)		$D = -0.69$ ($P > 0.1$)		$D = -1.64$ ($0.1 > P > 0.05$)	
Alu-rich subregion	0.182	0.202	0.145	0.135	0.163	0.135	0.182	0.213
(3820–4901)	0.150–0.214	0.026–0.378	0.107–0.183	0.003–0.267	0.129–0.197	0.003–0.267	0.164–0.200	0.057–0.369
	$D = -0.30$ ($P > 0.1$)		$D = 0.21$ ($P > 0.1$)		$D = 0.56$ ($P > 0.1$)		$D = -0.39$ ($P > 0.1$)	
3′ subregion	0.157	0.199	0.141	0.100	0.141	0.085	0.179	0.169
(4902–6600)	0.123–0.191	0.045–0.353	0.107–0.175	0.008–0.192	0.105–0.177	0.001–0.169	0.167–0.191	0.053–0.285
w/o *Alu*	$D = -0.68$ ($P > 0.1$)		$D = 1.17$ ($P > 0.1$)		$D = 1.78$ ($0.1 > P > 0.05$)		$D = 0.17$ ($P > 0.05$)	
Entire region	0.136	0.209	0.116	0.114	0.122	0.110	0.141	0.211
	0.114–0.158	0.075–0.343	0.094–0.138	0.036–0.192	0.104–0.140	0.034–0.186	0.131–0.151	0.101–0.321
	$D = -1.20$ ($0.1 > P > 0.05$)		$D = 0.08$ ($P > 0.1$)		$D = 0.44$ ($P > 0.1$)		$D = -1.04$ ($P > 0.1$)	
Entire region	0.112	0.185	0.096	0.083	0.104	0.083	0.120	0.175
w/o *Alu*	0.92–0.132	0.063–0.307	0.078–0.114	0.021–0.145	0.088–0.120	0.078–0.114	0.112–0.128	0.079–0.271
	$D = -1.40$ ($P > 0.1$)		$D = 0.51$ ($P > 0.1$)		$D = 0.84$ ($P > 0.1$)		$D = -1.00$ ($P > 0.1$)	

π and θ are in percentage per site. 95% confidence intervals are given below values; w/o, without.

[a] The numbers in parentheses refer to the positions in the sequenced region.

TABLE 3

Relative frequencies of sequence variants and subdivision among continents

Site	Variants[a]	Africans ($2n = 36$)	Asians ($2n = 36$)	Europeans ($2n = 36$)	χ^2	F_{ST}
457	G/A	0.167	0.583	0.167	19.6*	0.181
552[b]	G/A	0.167	0.139	0.056	2.3	0.021
564	C/T	0.083	0	0	6.1	0.057
631	G/C	0.056	0.056	0.111	1.1	0.010
665	T/C	0.028	0	0.083	3.6	0.033
1028	C/T	0.083	0	0	6.1	0.057
1051[b]	G/C	0.167	0	0	12.7	0.118
1266	C/T	0.417	0.611	0.361	5.0	0.046
1708	C/T	0.056	0.167	0.500	21.0*	0.195
2065	C/A	0.139	0.139	0.111	0.2	0.002
2173	A/G	0.056	0.167	0.528	23.4*	0.216
2372[b]	A/G	0.139	0.139	0.111	0.2	0.002
2973	G/A	0	0.056	0.028	2.1	0.019
3013[b]	G/A	0	0.139	0.083	5.1	0.048
3327	C/T	0.056	0.139	0.500	22.5*	0.209
3413[b]	G/A	0.417	0.222	0.139	7.6	0.071
4172	G/A	0	0	0.083	6.1	0.057
4206[b]	G/A	0.194	0.111	0	7.5	0.069
4288	G/A	0.250	0.667	0.194	20.6*	0.191
4333	T/C	0.194	0.083	0.111	2.1	0.020
4468[b]	T/C	0.389	0.306	0.583	6.0	0.055
4485[b]	A/G	0.056	0.139	0.556	27.5*	0.255
5021–5022[b]	AT/—	0.556	0.722	0.389	8.1	0.075
5305[b]	C/T	0.083	0.167	0.639	31.0*	0.287
5313	C/T	0.083	0	0	6.1	0.057
5539[b]	C/T	0.083	0.167	0.639	31.0*	0.287
5674	G/A	0.083	0.194	0.667	32.1*	0.297
6112[c]	C/T	0.611	0.583	0.278	9.8	0.091
6157[c]	G/A	0.444	0.583	0.167	13.5	0.125
6376[b,c]	A/T	0.611	0.667	0.167	21.6*	0.200
Average						0.112
Average with singletons and doubletons included						0.057

Singletons and doubletons were not included in this table. Only frequencies of the least frequent allele are shown. Statistical significance of the differences of allele frequencies among the three continents was examined with the χ^2 test (d.f. = 2). *Sites at which allele frequencies among continents are significantly different ($P <$ 0.001 derived from $P < 0.05$ using Bonferroni correction for 30 tests).

[a] The most frequent variant is shown first.

[b] Sites that potentially change their binding affinities to transcription factors.

[c] Sites located within the minimal promoter.

(0.209; 95% C.I. = 0.075–0.343) as in Asians (0.114; 95% C.I. = 0.036–0.192) or Europeans (0.110; 95% C.I. = 0.034–0.186); however, the differences again were not significant.

We examined differences among the continents at each polymorphic site (excluding singletons and doubletons; Table 3). Some variants were restricted to particular continents. Sites 564, 1028, 1051, and 5313 were polymorphic only in Africans, site 4172 was polymorphic only in Europeans, sites 2973 and 3013 were polymorphic only in Asians and Europeans, site 665 was polymorphic only in Africans and Europeans, and site 4206 was polymorphic only in Africans and Asians. However, the frequencies of the less common variant alleles at these sites were usually low (<0.1; the notable exceptions are sites 1051, 3013, and 4206). Also, the allele frequencies at individual sites were different among the continents. In fact, at 10 polymorphic sites the difference in allele frequencies among continents was statistically significant (Table 3). We calculated Wright's F_{ST} to measure the amount of differentiation among the continents (Table 3). The average F_{ST} for 76 polymorphic sites was 0.057 (it was 0.112 with singletons and doubletons excluded). F_{ST} at individual sites ranged from 0.002 to 0.297.

Distribution of diversity and divergence along the sequence: As an ~1-kb minisatellite-containing region and an ~0.3-kb region containing parts of *Alu* repeats were not sequenced, our sequencing resulted in three continuous fragments: 1–2488 bp, 2489–4362 bp, and

4363–6600 bp (Figure 1; we kept the numbering contiguous in spite of two gaps in the sequence). The subdivision of the sequence into the three fragments is based solely on our inability to sequence through a minisatellite (1 kb long) and the 0.3-kb region of *Alu*s. The *Alu*Y repeat was located in 433–714 bp, two *Alu*S repeats were located in 3820–4362 bp, and another two *Alu*S repeats were located in 4363–4901 bp. We analyzed *Alu*-containing regions separately because they might have higher values of nucleotide diversity compared to surrounding regions (Nachman and Crowell 2000; Chen and Li 2001). To investigate whether nucleotide diversity was evenly distributed along the sequence, estimates of π and θ were compared among the 5′ (1–432 bp and 715–2488 bp with *Alu*Y excluded), *Alu*Y subregion (433–714 bp), middle (2489–3819 bp), *Alu*-rich (3820–4901 bp), and 3′ (4902–660 bp) subregions (Table 2); the 3′ subregion is adjacent to the MC1R coding sequence (Figure 1). A sliding window analysis of π and θ provides a graphical representation of the results (Figure 2A). The average numbers of differences between any two sequences in a sample (π) were about half in the 5′ ($\pi = 0.094$; 95% C.I. = 0.082–0.106) and middle ($\pi = 0.086$; 95% C.I. = 0.070–0.102) subregions compared to the *Alu*-rich ($\pi = 0.182$; 95% C.I. = 0.164–0.200) and 3′ ($\pi = 0.179$; 95% C.I. = 0.167–0.191) subregions (Table 2). This difference was statistically significant and suggests a smaller number of high frequency variants in the 5′ and middle subregions compared to the *Alu*-rich and 3′ subregions. The *Alu*-rich subregion had values of nucleotide diversity ($\theta = 0.213$; 95% C.I. = 0.057–0.369; see π above) only slightly higher (and not significantly so) than that within the adjacent 3′ subregion ($\theta = 0.169$; 95% C.I. = 0.053–0.285). *Alu*Y had very high estimates of π (0.386; 95% C.I. = 0.318–0.454) and of θ (0.874; 95% C.I. = 0.248–1.500) compared to the other subregions analyzed (the difference for π is statistically significant, while it is not for θ). This repeat had 13 polymorphic sites within 280 bp, 9 of which were low frequency variants.

The distribution of nucleotide diversity along the sequence was different in Africans compared to Europeans and Asians (Table 2; Figure 2A). In Africans there was an excess of low frequency variants over high frequency variants (*D* was negative, though not significant) in all five subregions. In Asians and Europeans there was an excess of low frequency variants in the middle and *Alu*Y subregions (not significant), but there were more high frequency variants than low frequency ones in the other subregions (not significant; Table 2; Figure 2A).

We examined the spatial distribution of divergence on the basis of a comparison of human and chimpanzee and human and gorilla sequences (Figure 2B). The 5′, middle, *Alu*-rich, and 3′ subregions had similar divergence. The *Alu*Y region had elevated divergence between human and gorilla, but not between human and chimpanzee.

Tests of departure from neutrality: To examine whether variation in the sequenced region is compatible with neutral evolution, we used several tests. First, using the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987) we compared polymorphism and divergence (estimated between human and chimpanzee sequences) of the sequenced region with that of a 10-kb noncoding region on human chromosome 22 (Zhao *et al.* 2000). The sequences from all three continents were pooled for this test. The chromosome 22 region showed no significant deviation under several neutrality tests (Zhao *et al.* 2000), so presumably it evolves neutrally and can be used as an adequate comparison in a HKA test. The polymorphism and divergence were remarkably similar between the two regions (note that our region is only ⅔ in length of the region sequenced on chromosome 22), and consequently the test was not significant (Table 4). Similarly, the HKA tests were also performed separately for sequences from each continent. The results were again not significant (not shown).

Second, Tajima's test and Fu and Li's test with an outgroup were applied to the pooled sample and to each of the three continent samples (Table 5). When sequences from all three continents were pooled, Tajima's test was not significant, while Fu and Li's *D* and *F* were highly significant ($P < 0.005$). All three test statistics were negative for the pooled sample. Fu and Li's *D* and *F* were significantly negative for the pooled sample even when calculated with the *Alu*Y excluded ($D = -3.65$, $P < 0.02$; $F = -2.95$, $P < 0.02$), specifying a significantly high proportion of "young" *vs.* "old" mutations. When each continent was considered separately, Tajima's test was again not significant. However, Fu and Li's *F* and *D* were significantly negative in Africans ($P < 0.05$) and consistently positive (though not significant) in Asians and Europeans. This suggests that the sequenced region does not evolve neutrally in Africans. Furthermore, while Africans had a significantly high proportion of young mutations, Asians and Europeans had a high proportion of old mutations, although this was not significant.

The distribution of Fu and Li's *F*-statistic along the sequence is intriguing (Figure 3). For the pooled sample from three continents, *F* is negative in the 5′ and middle subregions of the sequence, but is around 0 in the 3′ subregion of the sequence (adjacent to the MC1R coding region). The distribution of *F*-statistic along the sequence in the African sample is similar to that of the pooled sample. In contrast, in both Asian and European samples *F* is positive or around zero for most of the sequence length, except for the middle subregion in Europeans.

We hypothesize the presence of selection in the middle subregion as it is conserved (compared to the 5′ and 3′ subregions) and is the only subregion to exhibit

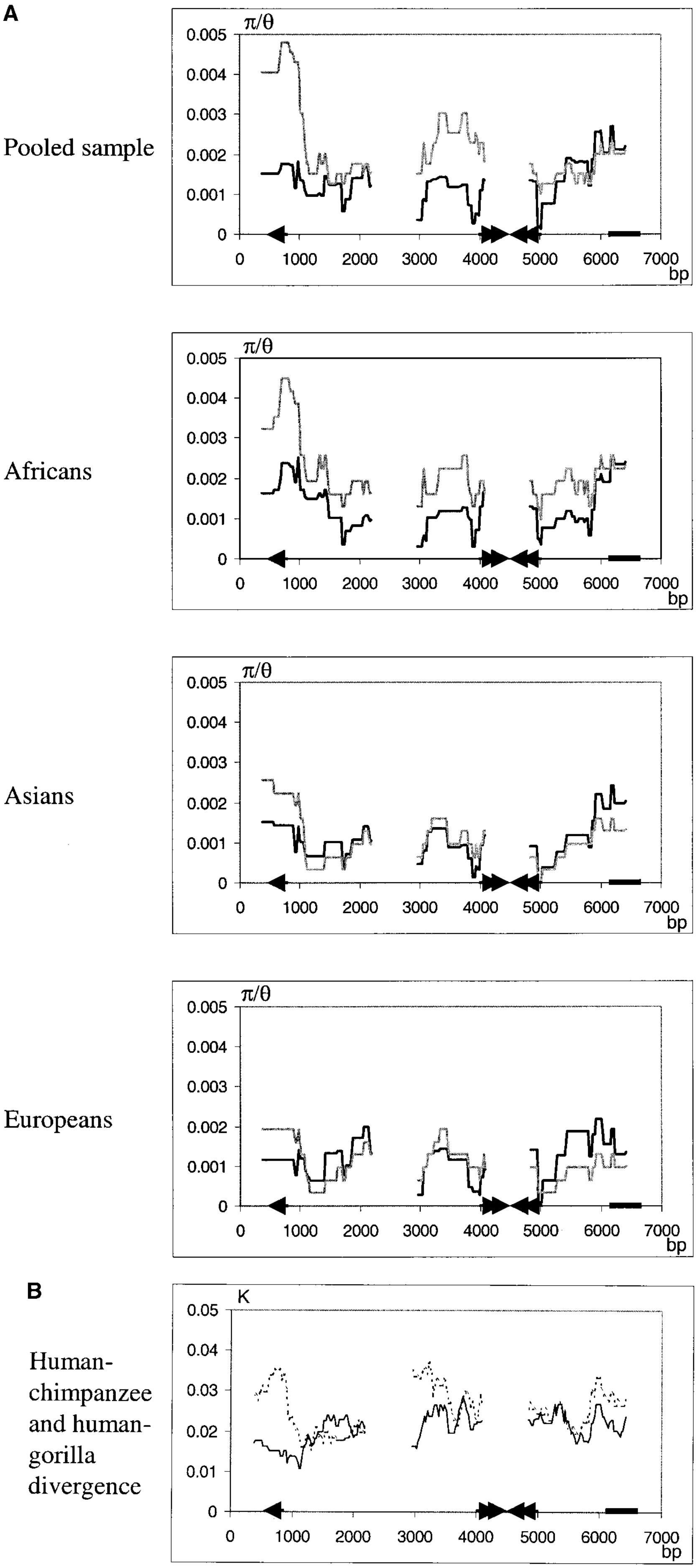

FIGURE 2.—Sliding window analysis of (A) nucleotide diversity (π in black and θ in gray) and (B) divergence between human and chimpanzee (solid line) and human and gorilla (dashed line). Minimal promoter and *Alu* repeats are indicated as in Figure 1.

TABLE 4

Results of the HKA test

	10 kb in 22q11.2	6.6 kb in 16q24.3
Intraspecific polymorphism data		
Segregating sites (obs)	71	72
Segregating sites (exp)	72.2	70.8
Total no. of sites	9901	6598
Sample size	128	108
Interspecific divergence		
No. differences (obs)	133.8	133.0
No. differences (exp)	132.6	134.2
$\chi^2 = 0.017$, $P = 0.8952$		

Indels were not considered.

a negative Tajima's D value in all three continents (see DISCUSSION). Additionally, Fu and Li's D is significantly negative for the African sequences and when all sequences are analyzed together (Table 6). To distinguish between purifying (or background) selection and directional selection (or hitchhiking) we compare Tajima's test statistic with Fu and Li's D statistic for this subregion (Table 6). Tajima's test statistic is less negative than Fu and Li's D for the pooled sample as well as for the African and European samples. This suggests that purifying but not directional selection is the more likely cause of the pattern observed in the middle subregion (FU 1997).

Mutation rate, parameter θ, effective population size, and age of the most recent common ancestor: The average numbers of nucleotide substitutions per site were 2.03% between the human and chimpanzee sequences, 2.59% between the human and gorilla sequences, and 5.62% between the human and orangutan sequences. The substitution rates were estimated to be 1.68×10^{-9}, 1.61×10^{-9}, and 2.02×10^{-9} per nucleotide per year by using a divergence time of 6 million years between human and chimpanzee, 8 million years between human and gorilla (Table 7), and 14 million years between human and orangutan (data not shown). Other divergence times were also considered (Table 7). The divergence times are based on estimates of GOODMAN *et al.* (1998) and from CHEN and LI (2001). Divergence data between human and orangutan were excluded from further analysis as we did not obtain a complete sequence for the region from orangutan and as we were mostly interested in parameter estimation for humans, so the most closely related outgroup species were chosen. The average mutation rate from comparisons between human and chimpanzee and between human and gorilla sequences is 1.65×10^{-9} per nucleotide per year, which is equal to 2.16×10^{-4} per sequence per generation.

Different methods were used to calculate the population parameter θ. It was estimated to be 8.64 by the average mutation rate per sequence per generation and an effective population size of 10,000 (TAKAHATA 1993), 13.70 by Watterson's method (WATTERSON 1975), and 9.14 by Tajima's method (TAJIMA 1983). The higher value of θ estimated by Watterson's method is due to an excess of singletons and doubletons. Watterson's and Tajima's θ-values were used to estimate the effective population size for several divergence times (Table 7). The results are largely in agreement with the commonly accepted estimate of 10,000.

TABLE 5

Neutrality tests

Test	No. of sequences	θ^{T}[a]	Test value	Critical value ($P = 0.05$)	Probability
African sequences					
Tajima's test	36	8.85	−1.195	−1.39	$P > 0.05$
Fu and Li's D	36	8.85	−1.945*	−1.92	$P < 0.05$
Fu and Li's F	36	8.85	−1.851*	−1.82	$P < 0.05$
Asian sequences					
Tajima's test	36	7.66	0.082	−1.41	$P > 0.10$
Fu and Li's D	36	7.66	0.122	−1.84	$P > 0.10$
Fu and Li's F	36	7.66	0.119	−1.75	$P > 0.10$
European sequences					
Tajima's test	36	7.92	0.441	−1.39	$P > 0.10$
Fu and Li's D	36	7.92	0.269	−1.84	$P > 0.10$
Fu and Li's F	36	7.92	0.369	−1.72	$P > 0.10$
All sequences					
Tajima's test	108	9.14	−1.040	−1.37	$P > 0.10$
Fu and Li's D	108	9.14	−4.044**	−1.88	$P < 0.005$
Fu and Li's F	108	9.14	−3.134**	−1.69	$P < 0.005$

* Significant at $P < 0.05$ level; ** significant at $P < 0.005$ level.

[a] θ^{T}, TAJIMA's (1983) θ.

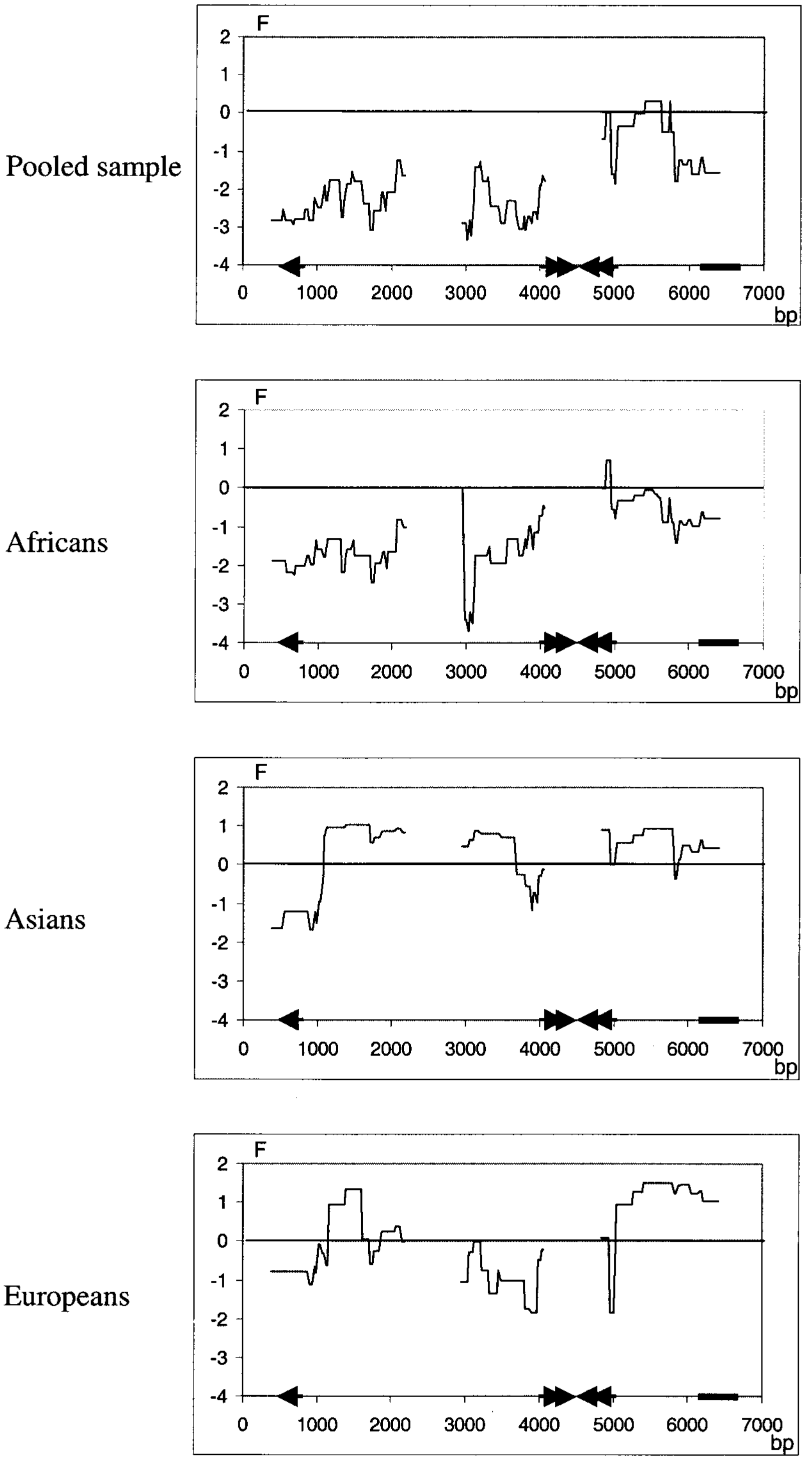

Figure 3.—Sliding window analysis of Fu and Li's (1993) *F*-statistic. Minimal promoter and *Alu* repeats are indicated as in Figure 1.

The age of MRCA was estimated for several values of effective population size for the entire sample, the African sample, and the non-African sample using the average mutation rate of 2.16×10^{-4} per sequence per generation (Table 8). Assuming the commonly used effective population size of 10,000 (Takahata 1993), the mode estimate (T_{mode}) and mean estimate (T_{mean}) are, respectively, 1,520,000 and 1,577,000 years for the entire sample. The estimates for the African sample are only slightly smaller than that for the entire sample, and the estimates for the non-Africans are about two to three times smaller than that for the entire sample, although these differences are not statistically significant (Table 8).

Polymorphism at transcription factor binding sites and phylogenetic footprinting: The potential effect of the variation at polymorphic sites on the function of the MC1R promoter was investigated. We examined

TABLE 6

Comparison between Tajima's statistic and Fu and Li's *D* in the middle subregion (2489–3819 bp)

Test	Statistic value	Probability
	African sequences	
Tajima's statistic	−1.72	$P > 0.05$
Fu and Li's *D*	−3.12*	$P < 0.05$
	Asian sequences	
Tajima's statistic	−0.42	$P > 0.10$
Fu and Li's *D*	0.22	$P > 0.10$
	European sequences	
Tajima's statistic	−0.69	$P > 0.10$
Fu and Li's *D*	−1.22	$P > 0.10$
	All sequences	
Tajima's statistic	−1.64	$P > 0.05$
Fu and Li's *D*	−4.47**	$P < 0.02$

* Significant at $P < 0.05$ level; ** significant at $P < 0.02$ level.

TABLE 8

The age (T, 10^3 yr) of the MRCA of human sequences

Sequences	N_e	T_{mode}	T_{mean}	95% interval
All samples	10,000	1,520	1,577	856–2,392
	12,000	1,325	1,412	739–2,227
	15,000	1,104	1,220	612–2,004
Africans	6,000	1,536	1,573	946–2,218
	8,000	1,382	1,473	838–2,182
	10,000	1,288	1,365	744–2,088
Non-Africans	6,000	768	811	389–1,344
	7,000	694	766	358–1,299
	8,000	646	728	339–1,248

The average mutation rate (2.16×10^{-4}/sequence/generation) was used.

whether any of the polymorphisms were located within the binding sites of transcription factors specified by Moro *et al.* (1999). In two alleles (belonging to a !Kung and a South African Bantu speaker) one of the SP-1 binding sites (sites 6002 and 6152, respectively) was disrupted; however, the second allele in each of these two individuals might compensate the promoter function. At the other transcription factor binding sites predicted by Moro *et al.* (1999) there was no variation. One of the TATA-boxes was not included in our analysis due to problematic sequencing in this part of the region. One of the E-boxes (site −2631 to −2626 according to Moro *et al.* 1999) was not confirmed by our sequencing.

We also tested whether changes at other sites might be important for MC1R promoter function. Potential transcription binding sites were predicted by comparison with the TRANSFAC database. The variation at 12 variant sites (Table 3) changes the recognition sites of the potential transcription factors. Notably, four of these sites (sites 4485, 5305, 5539, and 6376) had significantly different allele frequencies among the analyzed continents (Table 3).

Comparison of the transcription binding sites predicted by Moro *et al.* (2000) among human, chimpanzee, gorilla, and orangutan sequences revealed interesting features. The only E-box located within the minimal promoter in human had different copy number in different species. There was only one E-box at this position in human, but three in chimpanzee, and two in gorilla and orangutan (mouse also had two E-boxes at this site). In addition, two of the three experimentally proven SP-1 sites in human were disrupted in orangutan.

From ~800 bp of mouse MC1R promoter available (Adachi *et al.* 2000), 514 bp were aligned with the homologous segment from the human MC1R promoter with 58% similarity (Figure 4). Interestingly, the segment conserved between mouse and human sequences corresponds to the minimal promoter of the MC1R gene. The only E-box in the human MC1R minimal promoter is conserved between mouse and human. In mice, there were four more E-boxes within the minimal promoter compared to the human sequence. Four out of five polymorphic sites in the minimal promoter in human populations were conserved in human-mouse

TABLE 7

Comparison of the estimated values of mutation rate, parameter θ, and effective population size N_e

Parameters	Chimpanzee *vs.* human			Gorilla *vs.* human		
div. time, myr	5	6	7	7	8	9
ν (10^{-9}/site/year)	2.03	1.68	1.45	1.85	1.61	1.44
μ (10^{-4}/seq./gen.)	2.67	2.21	1.90	2.42	2.11	1.88
θ (N_e = 10,000)[a]	10.68	8.84	7.60	9.68	8.44	7.52
N_e (W)[b]	12,800	15,500	18,000	14,100	16,200	18,200
N_e (T)[c]	8,500	10,300	12,000	9,400	10,800	12,200

[a] θ was estimated by $\theta = 4N_e\mu$ using N_e = 10,000.

[b] N_e (W), the N_e value estimated by $\theta/4\mu$ using Watterson's (1975) θ-value (13.70).

[c] N_e (T), the N_e value estimated by $\theta/4\mu$ using Tajima's (1983) θ-value (9.14).

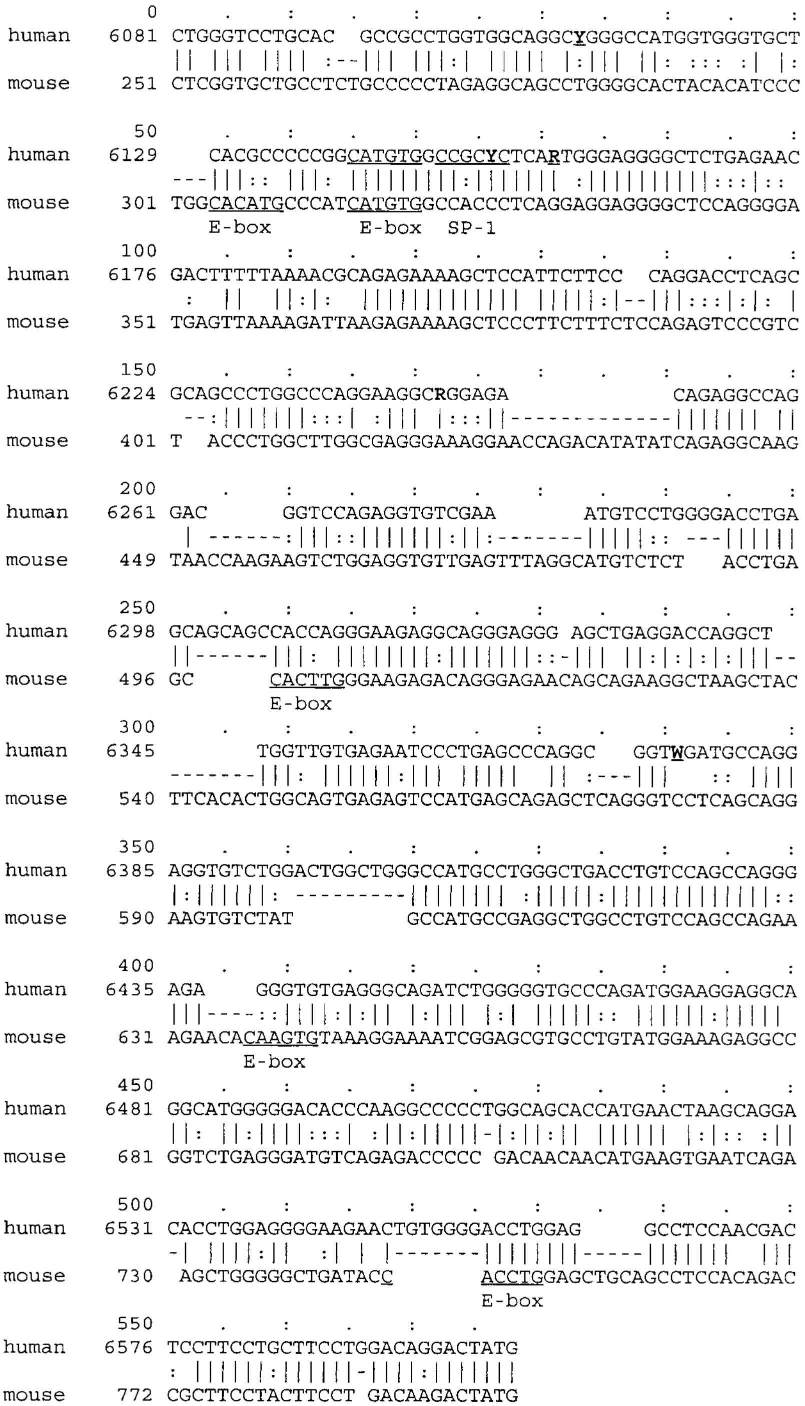
0
human 6081 CTGGGTCCTGCAC GCCGCCTGGTGGCAGGCYGGGCCATGGTGGGTGCT
mouse 251 CTCGGTGCTGCCTCTGCCCCCTAGAGGCAGCCTGGGGCACTACACATCCC
50
human 6129 CACGCCCCCGGCATGTGGCCGCYCTCARTGGGAGGGGCTCTGAGAAC
mouse 301 TGGCACATGCCCATCATGTGGCCACCCTCAGGAGGAGGGGCTCCAGGGGA
E-box
E-box
SP-1
100
human 6176 GACTTTTTAAAACGCAGAGAAAAGCTCCATTCTTCC CAGGACCTCAGC
mouse 351 TGAGTTAAAAGATTAAGAGAAAAGCTCCCTTCTTTCTCCAGAGTCCCGTC
150
human 6224 GCAGCCCTGGCCCAGGAAGGCRGGAGA CAGAGGCCAG
mouse 401 T ACCCTGGCTTGGCGAGGGAAAGGAACCAGACATATATCAGAGGCAAG
200
human 6261 GAC GGTCCAGAGGTGTCGAA ATGTCCTGGGGACCTGA
mouse 449 TAACCAAGAAGTCTGGAGGTGTTGAGTTTAGGCATGTCTCT ACCTGA
250
human 6298 GCAGCAGCCACCAGGGAAGAGGCAGGGAGGG AGCTGAGGACCAGGCT
mouse 496 GC CACTTGGGAAGAGACAGGGAGAACAGCAGAAGGCTAAGCTAC
E-box
300
human 6345 TGGTTGTGAGAATCCCTGAGCCCAGGC GGTWGATGCCAGG
mouse 540 TTCACACTGGCAGTGAGAGTCCATGAGCAGAGCTCAGGGTCCTCAGCAGG
350
human 6385 AGGTGTCTGGACTGGCTGGGCCATGCCTGGGCTGACCTGTCCAGCCAGGG
mouse 590 AAGTGTCTAT GCCATGCCGAGGCTGGCCTGTCCAGCCAGAA
400
human 6435 AGA GGGTGTGAGGGCAGATCTGGGGGTGCCCAGATGGAAGGAGGCA
mouse 631 AGAACACAAGTGTAAAGGAAAATCGGAGCGTGCCTGTATGGAAAGAGGCC
E-box
450
human 6481 GGCATGGGGGACACCCAAGGCCCCCTGGCAGCACCATGAACTAAGCAGGA
mouse 681 GGTCTGAGGGATGTCAGAGACCCCC GACAACAACATGAAGTGAATCAGA
500
human 6531 CACCTGGAGGGGAAGAACTGTGGGGACCTGGAG GCCTCCAACGAC
mouse 730 AGCTGGGGGCTGATACC ACCTGGAGCTGCAGCCTCCACAGAC
E-box
550
human 6576 TCCTTCCTGCTTCCTGGACAGGACTATG
mouse 772 CGCTTCCTACTTCCT GACAAGACTATG

FIGURE 4.—Alignment of the human and mouse MC1R minimal promoter. Consensus binding sites of transcription factors are underlined. Polymorphic sites in humans are shown in boldface and high frequency sites are underlined in bold. The last three nucleotides constitute the start codon of the MC1R gene. The numbering of the mouse sequence is according to its GenBank entry (accession no. AF176016).

comparison: the consensus nucleotides at these sites in humans were the same as those in mouse.

DISCUSSION

High polymorphism in the sequenced region: The region sequenced in this study is more polymorphic than the other three regions (each ~10 kb long) studied with a similar sampling scheme—one on chromosome 1 (Yu *et al.* 2001), one on chromosome 22 (Zhao *et al.* 2000), and one on chromosome X (Kaessmann *et al.* 1999). There were more polymorphic sites in the sequenced 6.6-kb region than in any of the three 10-kb regions (Table 1). The number of variable sites and consequently nucleotide diversity (especially θ) are expected to increase with sample size. Our study examined a slightly smaller sample size compared to the ones for the two other autosomal regions and still showed a higher polymorphism. The study on chromosome X (Kaessmann *et al.* 1999) examined a smaller sample size than we did and its low polymorphism may be due in part to this and in part to a smaller effective population size for the X chromosome compared to an autosome. The average nucleotide diversity (π) in the present region (0.141%) is higher than that in any of the three 10-kb regions (Table 1) and than the average value across 16 loci (0.081%) reported by Przeworski *et al.* (2000). In fact, only 3 out of the 16 loci (LPL, β-globin, and PDHA1) had a nucleotide diversity value higher than that for the present region. The θ-value in the present region (0.211% or 0.175% with *Alu*s excluded) is higher than that in any of the 16 loci surveyed by Przeworski *et al.* (2000), even though for many of them a larger sample size was examined.

The high polymorphism in the present region may be due to a high mutation rate (see below), a high recombination rate, and the presence of *Alu* repeats. The studied region is located on the very tip of the long arm of chromosome 16 (16q24.3), and the local recombination rate there is ~3.76 cM/Mb (reported for a marker D16S3037, located about 10 cM from the MC1R gene; Payseur and Nachman 2000), which is higher than the average for the human genome (~1.5 cM/Mb; Payseur and Nachman 2000). The local recombination rate in the present region is also higher than that in the 10-kb region on chromosome 22 (~1.86 cM/Mb; Payseur and Nachman 2000) and the 10-kb region on chromosome X (~0.16 cM/Mb; Kaessmann *et al.* 1999). A positive correlation was found between nucleotide diversity and recombination rate in Drosophila and humans and was attributed to the regions with low recombination rates being subject to stronger background selection and/or selective sweeps (Begun and Aquadro 1992; Charlesworth 1994; Nachman *et al.* 1998).

The presence of *Alu* repeats contributed to the high level of variation in the sequenced region. The π-value is reduced from 0.141 to 0.120% and the θ-value is reduced from 0.211 to 0.175%, when *Alu*s are excluded from the analysis. High polymorphism at *Alu*s is explained by a prevalence of highly mutable CpG dinucleotides in these repeats, especially in young *Alu*s, such as *AluY* in the present region (Schmid 1998).

Nonuniform distribution of variants among continents: Comparison of the patterns of sequence variation among the three continents in the present region and other regions studied with a similar sampling scheme (Table 1; Kaessmann *et al.* 1999; Zhao *et al.* 2000; Yu *et al.* 2001) led to several conclusions. In all four regions there were more variants in the African sample than in the non-African sample. As in the other three regions (reviewed in Yu *et al.* 2001), in the present region the estimate of π was higher in Africans than in Asians or Europeans. The distribution of singletons, doubletons (low frequency variants), and other (high frequency) variants among the continents was investigated. The region sequenced here follows a pattern similar to that observed in the 10-kb region on 22q11.2 (Zhao *et al.* 2000): there were about equal numbers of low and high frequency variants in Africans and an excess of high frequency variants in non-Africans (Table 1). In the 10-kb region on chromosome X there was an excess of low frequency variants in Africans and an approximately equal number of low and high frequency variants in non-Africans. On the other hand, in the 10-kb region in chromosome 1 (Yu *et al.* 2001) there was a deficiency of low frequency variants in Africans and an excess of low frequency variants in non-Africans. This suggests that considerable variation exists in the intercontinental distribution of polymorphic sites among different regions of the genome. Future studies will determine whether the pattern observed in the present region and in the 10-kb region on chromosome 22 (Zhao *et al.* 2000) is the prevailing one.

In the present region there was a high range of F_{ST} values among sites. Some site variants were unique to particular continents (Table 3), but the frequencies of such variants were usually low. Importantly, allele frequencies at several sites were significantly different among continents (Table 3).

Nonneutral evolution and forces shaping the variation: Our analyses indicate that the present region has not evolved according to the neutral Wright-Fisher model. Highly significant negative F and D values (implying an excess of rare variants) for the pooled sample suggest purifying (or background) selection, directional selection (or selective sweep), or population growth. Pooling of data from different populations may result in a higher level of population subdivision. However, population subdivision tends to reduce rather than increase the proportion of low-frequency variants (Yu *et al.* 2001). Of course, in those cases where only two genes (one individual) were sampled from a subpopulation, the number of singletons may be increased, but this may be partly compensated by the above tendency.

The distributions of π, θ, and Fu and Li's F-statistic (Figure 2A; Figure 3) were not uniform among the continents. This implies either different demographic histories or different selective pressures among the continents. The pattern of distribution of Fu and Li's F-statistic in this 6.6-kb region among three continents is in sharp contrast to the uniform distribution of F in the APOE region among four populations (Fullerton *et al.* 2000). The African sample exhibited a significant excess of young mutations, when all the subregions were considered together. This can be explained by strong selective pressure specifically on Africans or population expansion in Africans or both. The latter possibility has been considered by many authors (reviewed in Przeworski *et al.* 2000). Also, an excess of high over low frequency variants (not significant) in Asians and Europeans suggests a phase of population size reduction in non-Africans, as proposed by Przeworski *et al.* (2000). The sequenced locus follows the pattern observed in six out of eight loci, analyzed in Przeworski *et al.*'s (2000) Table 3; Tajima's D is higher in non-Africans compared to Africans. As this is observed in the majority of the loci studied, it probably reflects a demographic effect.

The distributions of π, θ, and Fu and Li's *F*-statistic (Figure 2A; Figure 3) were also not uniform throughout the sequenced region. This observation can be explained by differential selective pressures acting on different parts of the sequence (or their different proximity to genes under selection). In particular, the 5′ and middle subregions were significantly less variable than the 3′ subregion. This suggests either that the 5′ and middle subregions evolve under functional constraints or that the 3′ region is under diversifying selection. The distribution of variation is also nonuniform among the three continents. In Africans the 5′ and middle subregions may be evolving under purifying selection. This hypothesis is supported by significantly negative Fu and Li's *D* and *F* (Table 6) in the middle subregion and by marginally significant Fu and Li's *F* in the 5′ subregion ($F = -2.15$, $0.05 < P < 0.10$). In Asians and Europeans part of the 5′ subregion close to the KIAA1049 gene and parts of the middle subregion may also be evolving under purifying or background selection (suggested by negative *F* values; Figure 3). We speculate that an unidentified part of the MC1R promoter, a silencing element that specifies a tissue-specific expression of the MC1R gene, might be located within the 5′ or middle region and this important element may be under functional constraints. On the other hand, low polymorphism in the 5′ subregion might be explained by its proximity to the KIAA1049 gene.

Positive (though nonsignificant) *F* and *D* values in Asians and Europeans (Figure 3) in the 3′ subregion and parts of the 5′ subregion as well as a high average nucleotide diversity (π) in the 3′ subregion suggest that some sites in these subregions may be evolving under diversifying selection (possibly including the ones involved in the MC1R promoter function in Asia and Europe) and/or that the 3′ subregion is linked to a gene under diversifying selection (Rana *et al.* 1999). The contrast between *F* values in Asians or Europeans and those in Africans is interesting: the values are negative in Africans (except for the small segment in the 3′ subregion), whereas they are mostly positive in Asians and Europeans. However, the test statistics were not significant, so the variation may be neutral. The suggestion that the MC1R gene is under relaxation of selective constraints in Asians and Europeans (Harding *et al.* 2000) is not supported by the present data, because European and Asian sequences are less polymorphic than African sequences in the 3′ subregion, contrary to the pattern observed in the coding region. However, the difference is not statistically significant and the hypothesis of relaxation of selective constraints in Asians and Europeans cannot be dismissed.

Although it is difficult to make firm conclusions because both demographic factors and selection left their signatures in shaping the genetic variation of the same region, the observed pattern of polymorphism is consistent with a population expansion in Africans. We also speculate about (1) a possible phase of population size reduction in non-Africans and (2) possible purifying selection in the 5′ and middle subregions. The hypothesis of diversifying selection acting on some sites in the 3′ subregion or perhaps on the MC1R coding sequence in Asians and Europeans requires further investigation. The mostly uniform distribution of divergence in human-chimpanzee and human-gorilla comparisons suggests that the evolutionary events shaping this region are recent (*i.e.*, they took place after the human-chimpanzee divergence).

Population parameters and age of the most recent common ancestor: The mutation rate within the sequenced region estimated from the comparison of human sequences with the chimpanzee and gorilla sequences (1.65×10^{-9} per site per year) is higher than either of the estimates obtained for the 10-kb regions on chromosome 1 (0.74×10^{-9}; Yu *et al.* 2001) and chromosome 22 (1.15×10^{-9}; Zhao *et al.* 2000). This correlates with the high nucleotide diversity and suggests the highest neutral evolutionary rate for the present region among the three regions compared. The effective population size estimated for several divergence times and various estimates of θ supports the commonly used value of 10,000. This is consistent with the estimates from the 10-kb regions on chromosome 1 (Yu *et al.* 2001) and chromosome 22 (Zhao *et al.* 2000). As the assumption of neutrality is most likely violated for the present region, the estimates of mutation rate and population parameters should be treated cautiously, at least in Africans.

The estimates of the age of MRCA for the entire sample, Africans, and non-Africans were similar to the previous estimates from the 10-kb regions on chromosome 1 (Yu *et al.* 2001) and chromosome 22 (Zhao *et al.* 2000). The results imply ancient genetic histories for Africans and for non-Africans. As pointed out by Yu *et al.* (2001), the age of MRCA might be overestimated because the method used (Fu 1996; Fu and Li 1996, 1997) does not correct for excess of singletons in the data set.

Sites with potential regulatory function and phylogenetic footprinting: Studies comparing levels of MC1R expression among humans of different continents are not available, but we may hypothesize different levels of MC1R expression among humans with different skin colors. Our study suggests polymorphic sites that may be important for differential MC1R expression and, as a consequence, pigmentation variation in humans. Promoter assays can be used to determine the role of mutations at these sites on the regulation of MC1R expression. Sites that have different allele frequencies among continents and polymorphisms that change recognition sites of potential transcription factors (Table 3) should be examined first. Examination of high frequency polymorphic sites within the minimal promoter may be fruitful, in view of the fact that most high frequency variant sites are conserved in mouse (*i.e.*, they may be functionally important) and may be subject to diversifying selection according to our neutrality tests. In addition, since SP-1 is a transcription factor that determines the basal level of expression (Bentley *et al.* 1994), the level of expression may be different for alleles with nonconsensus SP-1 sequences (two singleton alleles in Africans). Although these alleles were each present in only one chromosome from a total of 108 chromosomes examined, their frequency may be higher in the populations in which they were found (!Kung and South African Bantu speakers).

Phylogenetic footprinting of the MC1R promoter provides interesting insights. Different numbers of E-boxes within the minimal promoter may be an important regulatory mechanism of MC1R expression. It is known that MC1R is expressed at a level much lower in humans than in mice (~100-fold; Mountjoy 1994). This correlates with the number of E-boxes in the minimal promoter: only one in human, but five in mouse. Mouse mast cells were shown to bind a microphthalmia transcription factor, an important regulator of gene expression in skin, at two out of five E-boxes (Adachi *et al.* 2000)—at a site with the variable number of E-boxes in primates. Consequently, from the examined primate outgroups, chimpanzee (three E-boxes) may have the highest level of MC1R expression and gorilla and orangutan (two E-boxes) may be intermediate. Additionally, disruption of two out of three SP-1 sites in the orangutan might have weakened the orangutan promoter. This might contribute to the low eumelanin level in orangutan and, as a result, to the red color of its hair. This hypothesis should be tested in functional studies.

We are grateful to two anonymous reviewers for valuable suggestions, to Michael Jensen-Seaman for critical reading of the manuscript, to Nathaniel Pearson for assistance, and to Lynn Jorde, Pekka Pamila, and Laszlo Patthy for DNA samples. This work was supported by National Institutes of Health grants GM55759, GM30998, and HD38287.

LITERATURE CITED

Abdel-Malek, Z., I. Suzuki, A. Tada, S. Im and C. Akcali, 1999 The melanin-1 receptor and human pigmentation. Ann. NY Acad. Sci. **885:** 117–133.

Adachi, S., E. Morii, D.-K. Kim, H. Ogihara, T. Jippo *et al.*, 2000 Involvement of mi-transcription factor in expression of alpha-melanocyte-stimulating hormone receptor in cultured mast cells of mice. J. Immunol. **164:** 855–860.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature **356:** 519–520.

Bentley, N. J., T. Eisen and C. R. Goding, 1994 Melanocyte-specific expression of the human tyrosinase promoter: activation by the microphthalmia gene and role of the initiator. Mol. Cell. Biol. **14:** 7996–8006.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between human and other hominoids and the effective population size of the common ancestor of human and chimpanzee. Am. J. Hum. Genet. **68:** 444–456.

Don, R. H., P. T. Cox, B. J. Wainwright, K. Baker and J. S. Mattick, 1991 "Touch-down" PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res. **19:** 4008.

Fu, Y.-X., 1996 Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics **144:** 829–838.

Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Fu, Y.-X., and W.-H. Li, 1996 Estimating the age of the common ancestor of men from the *ZFY* intron. Science **272:** 1356–1357.

Fu, Y.-X., and W.-H. Li, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. **14:** 195–199.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider *et al.*, 1998 Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9:** 585–598.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66:** 1666–1679.

Harding, R. M., E. Healey, A. J. Ray, N. S. Ellis, N. Flanagan *et al.*, 2000 Evidence for variable selective pressures at MC1R. Am. J. Hum. Genet. **66:** 1351–1361.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

John, P., and M. Ramsay, 2000 MC1R gene variation in normally pigmented Southern African individuals. Am. J. Hum. Genet. **67** (Suppl. 2): 237.

Kaessmann, H., F. Heissig, A. von Haeseler and S. Paabo, 1999 DNA sequence variation in a noncoding region of low recombination on the human X chromosome. Nat. Genet. **22:** 78–81.

Kikuno, R., T. Nagase, K. Ishikawa, M. Hirosawa, N. Miyajima *et al.*, 1999 Prediction of the coding sequences of unidentified human genes. XIV. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res. **6:** 197–205.

Moro, O., R. Ideta and O. Ifuku, 1999 Characterization of the promoter region of the human melanocortin-1 receptor (MC1R) gene. Biochem. Biophys. Res. Commun. **262:** 452–460.

Mountjoy, K. G., 1994 The human melanocyte stimulating hormone receptor has evolved to become "super-sensitive" to melanocortin peptides. Mol. Cell. Endocrinol. **102:** R7–R11.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.*, 1998 DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. **19:** 233–240.

Nickerson, D. A., S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark *et al.*, 2000 Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res. **10:** 1532–1545.

Payseur, B. A., and M. W. Nachman, 2000 Microsatellite variation and recombination rate in the human genome. Genetics **156:** 1285–1298.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Rana, B. K., D. Hewett-Emmett, L. Jin, B. H.-J. Chang, N. Sambuughin *et al.*, 1999 High polymorphism at the human melanocortin 1 receptor locus. Genetics **151:** 1547–1557.

Rieder, M. J., S. L. Taylor, A. G. Clark and D. A. Nickerson, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59–62.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analyses. Bioinformatics **15:** 174–175.

Schmid, C. W., 1998 Does SINE evolution preclude *Alu* function? Nucleic Acids Res. **26:** 4541–4550.

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer *et al.*, 2000 PipMaker—a web server for aligning two genomic DNA sequences. Genome Res. **10:** 577–586.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Takahata, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. **10:** 2–22.

Valverde, P., E. Healey, I. Jackson, J. Rees and A. J. Thody, 1995 Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nat. Genet. **11:** 328–330.

Watterson, G. A., 1975 On the number of segregating sites. Theor. Popul. Biol. **7:** 256–276.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich *et al.*, 2000 TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28:** 316–319.

Xia, X., 2000 *Data Analysis in Molecular Biology and Evolution*. Kluwer Academic Publishers. Boston/Dordrecht/London.

Yu, N., Z. Zhao, Y.-X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

Zhao, Z., L. Jin, Y.-X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: Y.-X. Fu

APPENDIX

Primer sequences

Primer	Sequence (5′–3′)	PCR[a]
111F	CACTGTTTCTCCTATAAATGTAAATGGGTCAC	1410R
575F	GACAAGAGTCTCACTGTGTCGC	
957F	GCCCATGTAGCAAAGATCAGG	
524R	GTCAGATTCAACAGATAGTGGCATC	
918R	GGCACTTCTCTGCAAAACATGCT	
1410R	CCAGGAACTGCCAAAAGGATGAACTC	111F
1365F	CCCATCACTGTGTAATCGTCTAACCTG	2629R
1796F	GGGATCTGCACTCATCTCCAGG	
2185F	GCTGAGCCTACTTCCAATGAC	
1818R	GGGTTATCTCCCAACCATCTTC	
2235R	CCACAATCATGGCAGAGGCTAC	
2629R	CGAGGGCTGCGAGAGGTAAAAC	1365F
3770F	GCCCTGGATGCCAGACACTGTAT	5090R
4212F	CCCAGTTCTCATGCCCTTTCAAGT	
4685F	GCGTGTGTGAACAGAAACAGG	
4248R	ACCCCAGCCTCCACTGCTACC	
4707R	CAAACCATCTTCAAATCGGCAG	
5090R	GCCCTAAAATGTTTTAATTGAGGTACAACATA	3770F
211F	GCTTATGTGGCTGGTTCAGGTCTGTCATCC	1682R
627F	GGCTCATCCCTGTAATCTCAGCAT	
780F	AGCTAGTTGGGAGGCTGAGGCATAAGTATTG	
953F	TTTTGAGACTGTATCTCTGTTT	
536R	GGCTGGAGTGCAGTGGCATGATCTTGG	
846R	CGCAGGCTGGAGTGTAGTGGTGCATTC	
1218R	AAATTATTATTGCAGGGCACAG	
1601R	GTGAGGACATTAATATTTTTCATA	
1682R	GATGTTCGAGTTAAAATCCATCCTGTCTCTCGC	211F
1623F	ACACACGGAGGTGGCTTGTGAGTGGT	IN
1988F	GCGAGAGGTCTGCCTTTGATGTGG	
2429F	CCTGGTCCAGCCCCCAAATCTGC	
2863F	GACGGTCCAGAGGTGTCGAAATGTCC	
2001R	GCCCTGCACCAACAGCCACATCAAAG	
2383R	GTTCTGGAAACTGAGTGAGCCCTGC	
2802R	CGCTGAGGTCCTGGGAAGAATGGAG	
IN	GGTGGAGTTGAGGGAGCCCAGAAGTCTT	1623F

[a] The primer used for PCR with this primer.